1

BOUNDARY CROSSING PROBABILITY COMPUTATIONS IN THE ANALYSIS OF SCAN STATISTICS

Hock Peng Chan, I-Ping Tu, and Nancy Ruonan Zhang

National University of Singapore
Academia Sinica
Stanford University

Abstract: The theory of boundary crossing probabilities in the study of repeated likelihood ratio tests was developed by Lai, Siegmund and Woodroofe in a series of articles and monographs appearing in the late seventies and early to mid eighties. This form part of the foundation for subsequent developments in the analysis of maxima of Gaussian and Poisson random fields used to provide accurate tail probability approximations of scan statistics. In this paper, we (i) track these theoretical developments, (ii) study their applications on spatial scan statistics in astronomy and epidemiological studies and (iii) relate these theoretical developments to scan statistics used recently in genomics.

Keywords and phrases: Astronomy, boundary crossing probability, DNA copy number, epidemiology, genomics, maxima of random fields, neuroscience, scan statistic

1.1 Introduction

The study of scan statistics to detect either a signal at an unknown location or the presence of spatial clustering in a compact domain is a very active area of research and the areas of applications are diverse, including astronomy, epidemiology, genomics, neuroscience, botany and ecology. The basic idea is as follows. A list of spatial or space-time vectors $\mathbf{x}_1, \ldots, \mathbf{x}_J$ associated with the occurrence of certain events of interest are observed in a domain D. In addition, there may also be a random variable or vector X_j that provides additional information on the jth occurrence for each $1 \leq j \leq J$. If there is a source of a cluster at an unknown location $\mathbf{t}$ (or a signal centered at $\mathbf{t}$), it may result either in an unusually large number of occurrences near $\mathbf{t}$ or the distribution of

X_j might be different when $\mathbf{x}_j$ is near $\mathbf{t}$. For example, in case-control datasets in epidemiological studies, $X_j = 1$ denotes the occurrence of a case and $X_j = 0$ the occurrence of a control. When there is a source of a cluster of cases at $\mathbf{t}$, the probability that $X_j = 1$ will be higher when $\mathbf{x}_j$ is near $\mathbf{t}$. A score $S(\mathbf{t})$ is computed from $\{(\mathbf{x}_j, X_j) : 1 \le j \le J\}$ and a high score is expected when the source of the cluster is at $\mathbf{t}$. Since $\mathbf{t}$ is unknown, the scan statistic $M := \sup_{\mathbf{t}\in D} S(\mathbf{t})$ is the summary score for the presence of a cluster in D.

Lai and Siegmund (1977, 79), Woodroofe (1978, 82) and Siegmund (1985) developed a set of techniques to study boundary crossing probabilities of generalized likelihood ratio (GLR) sequential test statistics. These techniques were subsequently refined and extended by many researchers so that they can be applied on a wide variety of settings. We track these developments in Section 2 and elaborate on their applications in scan statistics in astronomy and epidemiology in Section 3 and genomics in Section 4. We conclude the paper with a few brief remarks in Section 5.

1.2 Theoretical Developments

Throughout this paper, $\mathbf{I}$ shall denote the indicator function, $|\cdot|$ the Lebesgue measure of a set or the determinant of a square matrix and $\|\cdot\|$ the L_2 norm. In addition, $\varphi(x) = (2\pi)^{-1/2}e^{-x^2/2}$ and $\Phi(y) = \int_{-\infty}^{y} \varphi(x)\, dx$ are the density and cumulative distribution respectively of the standard normal. We write $a_n \sim b_n$ if $\lim_{n\to\infty}(a_n/b_n) = 1$. If $\mathbf{t} = (t_1, \ldots, t_d) \in \mathbf{R}^d$ and A is a subset of $\mathbf{R}^d$, then for any $w > 0$, $\mathbf{t} + wA = \{\mathbf{t} + w\mathbf{u} : \mathbf{u} \in A\}$. Before proceeding to the analytical techniques, we give a few examples to illustrate how the scores $S(\mathbf{t})$ are defined in different settings.

Example 1. Let J be either a fixed positive integer or a Poisson random variable. Assume that under the null hypothesis of no clustering, $\mathbf{x}_1, \ldots, \mathbf{x}_J$ are independent and identically distributed (i.i.d.) random variables uniformly distributed on a compact domain D. Let A be a nice compact set, for example the box-kernel $A = \{\mathbf{u} : \max_i |u_i| \le w/2\}$ or the spherical kernel $A = \{\mathbf{u} : \|\mathbf{u}\| \le w\}$ for some $w > 0$. Let $S(\mathbf{t})$ be the number of occurrences $\mathbf{x}_j$ lying inside $\mathbf{t}+A$ and M the corresponding scan statistic. Naus (1965, 66, 82), Huntington and Naus (1975) and Glaz (1989) provided approximate and exact p-value calculations of M when A is the box-kernel. See Glaz, Naus and Wallenstein (2001) for comparisons against competing p-value approximations and bounds and also for a good overview of recent developments in scan statistics.

Example 2. Let $\mathbf{x}_1, \ldots, \mathbf{x}_J$ be the points on a lattice grid in a compact domain D. The detection of a signal is of interest here. Under the null hypothesis of no signal, $X_1, \ldots, X_J$ are i.i.d. random variables from a baseline

distribution F with log moment generating function $\psi(\theta) := \log Ee^{\theta X_1}$. Assume that $\Theta := \{\theta : \psi(\theta) < \infty\}$ is finite in a neighborhood of 0. Then the rate function of F is given by $\phi(\mu) = \sup_{\theta\in\Theta}[\theta\mu - \psi(\theta)]$ and F can be embedded in an exponential family $\{F_\theta, \theta \in \Theta\}$, with $dF_\theta(x) = e^{\theta x-\psi(\theta)}dF(x)$. Let A be a given signal shape and consider the alternative hypothesis

H_1: there exists $\theta \neq 0$ and $\mathbf{t} \in D$ such that $X_1, \ldots, X_J$ are independent with $X_j \sim F_\theta$ if $\mathbf{x}_j \in \mathbf{t} + A$ and $X_j \sim F$ otherwise,

indicating that a signal of shape A is centered at some unknown $\mathbf{t} \in D$. The log GLR score for testing the null hypothesis against the alternative hypothesis is $S(\mathbf{t}) = n_\mathbf{t}\phi(\bar{X}_\mathbf{t})$, where $n_\mathbf{t}$ is the number of points $\mathbf{x}_j$ lying in $\mathbf{t} + A$ and $\bar{X}_\mathbf{t} = n_\mathbf{t}^{-1}\sum_{\mathbf{x}_j\in\mathbf{t}+A} X_j$. Tail probabilities for the maxima of $S(\mathbf{t})$ were computed in Siegmund and Yakir (2000) via a change of measure argument.

Example 3. Researchers in neuroscience are interested to know if a neural spike time pattern, for example the pattern observed when a bird is learning a new song while awake, is repeated when the bird is sleeping, see Dave and Margoliasch (2000) for a more elaborate introduction to the problem. Let $T > 0$ and $\mathcal{Y} = \{y_1, \ldots, y_N\}$ a given template spike time pattern with $0 \le y_n \le T$ for all n and $\mathcal{X} = \{x_1, \ldots, x_J\}$ the neural spike times when the bird is sleeping, with $0 \le x_j \le U$ for all j, U large compared to T. We want to check if the spike-time pattern $\mathcal{Y}$ is repeated inside $\mathcal{X}$, in other words, if there exists a time t such that $t + \mathcal{Y}$ and $\mathcal{X} \cap [t, t + T]$ are similar.

In Chi, Rauske and Margoliasch (2003), a pattern-filtering algorithm was used to match the spike time patterns. Let f be a non-increasing kernel scoring function on $[0, \infty)$ with $f(0) > 0$ and $\lim_{u\to\infty} f(u) < 0$. Common examples include the continuous Hamming window kernel

$$f(u) = \begin{cases} \frac{1}{2}(1-\beta) + \frac{1}{2}(1+\beta)\cos\left(\frac{\pi u}{\epsilon}\right) & \text{if } u < \epsilon \\ -\beta & \text{if } u \ge \epsilon, \end{cases}$$

or the box-kernel

$$f(u) = \begin{cases} 1 & \text{if } u < \epsilon \\ -\beta & \text{if } u \ge \epsilon. \end{cases}$$

The score

$$S(t) = \sum_{x_j\in[t,t+T]} \max_{1\le n\le N} f(|x_j - t - y_n|)$$

provides the value of a match between $t + \mathcal{Y}$ and $\mathcal{X} \cap [t, t + T]$. In Chi (2004), under the assumption that x_1 and $x_{i+1} - x_i$, $i \ge 1$, are i.i.d. exponential random variables, the exponent of the tail probability of $M = \sup_t S(t)$ was obtained using large deviation theory. Using the theory of boundary crossing probabilities, Chan and Loh (2007) obtained a more precise estimate, an approximation of the tail probability of M.

We shall illustrate the techniques behind the computation of boundary crossing probabilities with the signal detection problem described in Example 2. Let $d = 1$ and $\bar{X}_{i,j} = (j-i)^{-1} \sum_{i+1}^{j} X_k$ when $i < j$. Let $X_1, \ldots, X_J$ be i.i.d. random variables with distribution F under the null hypothesis and let the score

$$S(i,j) = (j-i)\phi(\bar{X}_{i,j}),$$

where ϕ is defined in Example 2. Let the scan statistic

$$M = \sup_{0 \le i < j \le J, w_0 \le (j-i) \le w_1} S(i,j).$$

We shall consider here the computation of $P\{M \ge c\}$ when $\log J = o(c)$, $J/c \to \infty$ and $w_k \sim \alpha_k c$ for some $0 < \alpha_0 < \alpha_1$ as $c \to \infty$. The problem has applications in sequential change-point detection, and is solved for normal X_j when $w_0 = 0$ and $w_1 = \infty$, in Siegmund and Ventrakaman (1995), and extended to Markovian X_j satisfying minorization and drift conditions and ϕ replaced by a general function in Chan and Lai (2002, 2003).

Large deviation approximations. Let $v_\mu = \frac{d^2}{d\theta^2}\psi(\theta)|_{\theta=\theta_\mu}$ and $\Lambda = \{\mu : \alpha_1^{-1} \le \phi(\mu) \le \alpha_0^{-1}\}$. Assume for convenience that F has a continuous bounded density and Λ is a compact set lying in the interior of the support of F. Then the saddlepoint approximation

$$P\{\bar{X}_{i_0,j_0} \in d\mu\} \sim \left(\frac{j_0 - i_0}{2\pi v_\mu}\right)^{1/2} e^{-(j_0-i_0)\phi(\mu)}\, d\mu, \tag{1.1}$$

holds uniformly over $\mu \in \Lambda$. Our interest is focused on μ satisfying $(j_0 - i_0)\phi(\mu) = c + x$ for some x either of order 1 or small compared to c.

Local random walk. The next step involves an analysis of the local behavior of $S(i,j)$ for (i,j) close to (i_0, j_0) when $S(i_0, j_0) = c + x$. Let $\mu = \bar{X}_{i_0,j_0}$ and let $\theta_\mu \in \Theta$ satisfies $\phi(\mu) = \theta_\mu \mu - \psi(\theta_\mu)$. Since $\frac{d}{d\mu}\phi(\mu) = \theta_\mu$, it follows from a Taylor series expansion that

$$\begin{aligned} S(i,j) &= (j-i)\phi((\bar{X}_{i,j} - \mu) + \mu) \doteq (j-i)[\phi(\mu) + (\bar{X}_{i,j} - \mu)\theta_\mu] \\ &= S(i_0,j_0) + \sum_{k=1}^{J} (\mathbf{I}_{\{k \in [i,j]\}} - \mathbf{I}_{\{k \in [i_0,j_0]\}})[\theta_\mu X_k - \psi(\theta_\mu)]. \end{aligned} \tag{1.2}$$

Clearly X_k follows distribution F for $k \le i_0$ and $k > j_0$ irregardless of the conditioning on $\bar{X}_{i_0,j_0}$. In addition, by Siegmund (1988), X_k is asymptotically of distribution F_μ (that is F_{θ_μ}) and asymptotically independent (for a fixed number of random variables) for $i_0 < k \le j_0$, when we condition on $\bar{X}_{i_0,j_0} = \mu$. Hence under the conditioning,

$$\sum_{k=1}^{J} (\mathbf{I}_{\{k \in [i,j]\}} - \mathbf{I}_{\{k \in [i_0,j_0]\}})[\theta_\mu X_k - \psi(\theta_\mu)] \Rightarrow W_{i-i_0} + \widetilde{W}_{j-j_0}, \tag{1.3}$$

where W and $\widetilde{W}$ are independent random walks with independent increments $[\theta_\mu X_n - \psi(\theta_\mu)]$ and $[\theta_\mu \widetilde{X}_n - \psi(\theta_\mu)]$ respectively, with $X_n \sim F_\mu$ for $n \geq 1$, $X_n \sim F$ for $n \leq 0$, $\widetilde{X}_n \sim F$ for $n \geq 1$ and $\widetilde{X}_n \sim F_\mu$ for $n \leq 0$. We shall denote by P_μ the probability when W and $\widetilde{W}$ have increments with these joint distributions.

We are now left with the task of combining these large deviation approximations and local random walks and we shall highlight three approaches here.

(I) Conditioning on the last-exit (or first-passage) time. This is the method most closely identified with the techniques developed to analyze sequential GLR test statistics. Unlike in sequential analysis where only one index is involved and what the last time is needs no explanation, here we need to deal with two indices i and j. We handle this by defining an ordering $\succ$ with $(i,j) \succ (i_0,j_0)$ if either $i > i_0$ and $j = j_0$ both occurs or if $j > j_0$ occurs. By (1.1)-(1.3), if $(j_0 - i_0)\phi(\mu) = c + x$, then

$$\begin{aligned} &P\{\bar{X}_{i_0,j_0} \in d\mu, (j-i)\phi(\bar{X}_{i,j}) < c \text{ for all } (i,j) \succ (i_0,j_0)\} \\ &\sim \left(\frac{c+x}{2\pi\phi(\mu)v_\mu}\right)^{1/2} e^{-c-x} P_\mu\left\{\max_{k\geq 1} W_k \leq -x\right\} \\ &\quad \times P_\mu\left\{\max_{k\leq 0} W_k + \max_{\ell \geq 1} \widetilde{W}_\ell \leq -x\right\}\, d\mu. \end{aligned} \tag{1.4}$$

We sum (1.4) over $j_0 \geq i_0 + c/\phi(\mu)$ for a fixed i_0, noting that x increases by $\phi(\mu)$ for each increase of j_0 by 1, integrate over $\mu \in \Lambda$ and sum over $1 \leq i_0 \leq J$ to obtain

$$P\{M \geq c\} \sim J\left(\frac{c}{2\pi}\right)^{1/2} e^{-c} \int_\Lambda \gamma(\mu)(\phi(\mu))^{-3/2} v_\mu^{-1/2}\, d\mu, \tag{1.5}$$

where

$$\gamma(\mu) = \int_0^\infty e^{-x} P_\mu\left\{\max_{k\geq 1} W_k \leq -x\right\} P_\mu\left\{\max_{k\leq 0} W_k + \max_{\ell\geq 1} \widetilde{W}_\ell \leq -x\right\}\, dx.$$

A rigorous justification of (1.5) is more involved, as given in Siegmund and Venkatraman (1995) for the case of normal X_i. They also provided a simplification, relating γ to the overshoot constant

$$\nu(x) = 2x^{-2} \exp\Big\{-2\sum_{n=1}^\infty n^{-1/2}\Phi\Big(-\frac{x\sqrt{n}}{2}\Big)\Big\} \quad (x > 0), \tag{1.6}$$

in the normal case. This is achieved via an identity in Siegmund (1992). Analogous overshoot constant expressions for general X_i, relevant to both p-value and sample size calculations, can be found in Woodroofe (1979), Tu and Siegmund (1999), Storey and Siegmund (2001) and Tu (2008).

(II) Conditioning on local or global maxima. Let (i_0, j_0) be the indices at which the maximal value $M = S(i_0, j_0) \geq c$ is attained. By (1.1)-(1.3), we obtain (1.5) with the alternative representation

$$\gamma(\mu) = P_\mu \left\{ \max_{k \neq 0} W_k < 0 \right\} P_\mu \left\{ \max_{\ell \neq 0} \widetilde{W}_\ell < 0 \right\}.$$

This approach is more commonly used when the score is obtained via a continuous kernel function. A good reference is Rabinowitz and Siegmund (1997), which considers signal detection on a homogeneous Poisson process. This work is discussed in more detail in Section 3.1.

(III) Conditioning below a high crossing. The first two approaches involve conditioning above a high level c. There is yet another approach, adapted by Hogan and Siegmund (1986) from tail probability approximations of Gaussian random fields developed in Pickands (1969), Bickel and Rosenblatt (1973) and Qualls and Watanabe (1973). Fix i_0 and j_0 and let them be multiples of n for some large n. We condition on $S(i_0, j_0) < c$, compute the conditional probability that $S(i, j)$ exceeds c for some (i, j) lying in the domain $[i_0, i_0 + n] \times [j_0, j_0 + n]$, then add up these probabilities over different $i_0 < j_0$. By (1.1)-(1.3), if $(j_0 - i_0)\phi(\mu) = c - x$, then

$$\begin{aligned} &P\{\bar{X}_{i_0,j_0} \in d\mu, (j-i)\phi(\bar{X}_{i,j}) \geq c \text{ for some } (i,j) \in [i_0, i_0+n] \times [j_0, j_0+n]\} \\ &\quad \sim \left(\frac{c-x}{2\pi\phi(\mu) v_\mu} \right)^{1/2} e^{-c+x} P_\mu \left\{ \max_{0 \leq k \leq n} W_k + \max_{0 \leq \ell \leq n} \widetilde{W}_\ell \geq x \right\} d\mu. \end{aligned} \tag{1.7}$$

We sum (1.7) over $i_0 \leq j_0 \leq i_0 + c/\phi(\mu)$ with j_0 a multiple of n and i_0 fixed, integrate over $\mu \in \Lambda$, then sum over $1 \leq i_0 \leq J$ with i_0 a multiple of n, while choosing n large, to obtain (1.5) with

$$\gamma(\mu) = \lim_{n \to \infty} n^{-2} \int_{-\infty}^{\infty} e^x P_\mu \left\{ \max_{0 \leq k \leq n} W_k + \max_{0 \leq \ell \leq n} \widetilde{W}_\ell \geq x \right\} dx.$$

Again, additional technical arguments are needed here for a rigorous justification of these calculations. This approach was used in Chan and Zhang (2007) to compute tail probabilities of weighted scan statistics and in Chan and Loh (2007) to compute tail probabilities of template scoring scan statistics. The first problem will be elaborated further in Section 4.1.

1.3 Applications in spatial scan statistics

We focus here on two examples to illustrate how the theory of boundary crossing probabilities can be used to obtain analytical p-values for spatial or space-time

scan statistics. We start off on a problem with motivations in astronomy. Note that the calculations for continuous kernel functions [Rabinowitz and Siegmund (1997)] and kernels containing discontinuities [Loader (1991)] are different. We then consider the problem of detecting clusters in a nonhomogeneous population using a case-control dataset.

1.3.1 Searching for a source of muon particles in the sky

Continuous kernel functions. Consider a background of homogeneous random cosmic rays with known intensity λ. By taking D sufficiently large, we may assume that edge effects are absent and that the particles are observed on $\mathbf{R}^d$. We shall denote the set of observed particle locations by $\{\mathbf{x}_j\}_{j=1}^{\infty}$. Let f be a non-negative kernel function on $\mathbf{R}^d$ satisfying $\int f^2(\mathbf{x})d\mathbf{x} = 1$, is smooth and symmetric in each argument, and vanishes rapidly at infinity. One concrete example is the Gaussian kernel $f(\mathbf{x}) = \pi^{-d/4}e^{-\|\mathbf{x}\|^2/2}$. Let $\mu = \int f(\mathbf{x})d\mathbf{x}$ and let the score

$$S(\mathbf{t}) = \lambda^{-1/2}\Big[\sum_{j=1}^{\infty} f(\mathbf{x}_j - \mathbf{t}) - \lambda\mu\Big]. \tag{1.8}$$

Let $P_{\theta,\mathbf{t}}$ ($E_{\theta,\mathbf{t}}$) denote the probability measure (expectation) under which $\{\mathbf{x}_j\}_{j=1}^{\infty}$ is generated from a nonhomogeneous Poisson process with intensity

$$\lambda_{\theta,\mathbf{t}}(\mathbf{x}) := \lambda \exp[\theta f(\mathbf{x} - \mathbf{t})], \tag{1.9}$$

and let $P_{\theta,\mathbf{0}}$ be denoted more simply by P_θ. The nonhomogeneous Poisson process motivates $S(\mathbf{t})$ as the efficient score statistics as we let $\theta \to 0$ and also provides the change of measure for computing the tail probabilities of the scan statistic $M = \sup_{t\in D} S(\mathbf{t})$.

We provide an outline of the calculations and arguments given in Rabinowitz and Siegmund (1997) and refer the reader to the article itself for the details. Fix $c > 0$ and let $b = c\lambda^{1/2}$. By the Poisson clumping heuristic, see for example Siegmund (1988) or Aldous (1989),

$$P_0\{M \geq b\} \approx 1 - e^{-E_0 K},$$

where K is the number of local maxima in D exceeding the threshold b. Since f is smooth, $\nabla S(\mathbf{t})$ and $\nabla^2 S(\mathbf{t})$, the gradient and Hessian respectively of S at $\mathbf{t}$, are both well-defined and continuous. It follows from Theorem 6.1 of Adler (1981), using a local maxima conditioning argument, that

$$E_0 K = |D| E_\theta \left[\left(\frac{dP_0}{dP_\theta}\right) |\nabla^2 S(\mathbf{0})| \mathbf{I}_{\{S(\mathbf{0})\geq b, \nabla S(\mathbf{0})=0, \nabla^2 S(\mathbf{0})<0\}}\right], \tag{1.10}$$

where the statement "$\nabla^2 S(\mathbf{0}) < 0$" means $\nabla^2 S(\mathbf{0})$ is a negative definite matrix, and the expectation on the right hand side of (1.10) is defined with respect to

a joint probability-density. Let

$$\psi(\theta) = \log E_0[e^{\theta\lambda^{1/2}S(\mathbf{0})}] = \lambda \int [e^{\theta f(\mathbf{x})} - 1 - \theta f(\mathbf{x})]\, d\mathbf{x}.$$

Then

$$\begin{aligned} E_\theta(\lambda^{1/2}S(\mathbf{0})) &= \psi'(\theta) = \lambda \int f(\mathbf{x})[e^{\theta f(\mathbf{x})} - 1]d\mathbf{x}, \\ \mathrm{Var}_\theta(\lambda^{1/2}S(\mathbf{0})) &= \psi''(\theta) = \lambda \int f^2(\mathbf{x})e^{\theta f(\mathbf{x})}d\mathbf{x}. \end{aligned}$$

Let the rate function $I(\theta) = \theta\psi'(\theta) - \psi(\theta)$ and select θ to satisfy $\psi'(\theta) = c\lambda$. By a Gaussian approximation on the process $S(\mathbf{t})$ under P_θ, and making use of the relations

$$\begin{aligned} &E_\theta[\nabla S(\mathbf{0})] = 0, \qquad \mathrm{Cov}_\theta(S(\mathbf{0}), \nabla S(\mathbf{0})) = \mathbf{0}, \\ &E_\theta[\lambda^{1/2}\nabla^2 S(\mathbf{0})] = -\theta \mathrm{Cov}_\theta(\lambda^{1/2}\nabla S(\mathbf{0})), \qquad E_\theta(\nabla^2 S(\mathbf{0}), \nabla S(\mathbf{0})) = \mathbf{0}, \\ &\mathrm{Cov}_\theta\left(\frac{\partial}{\partial t_i}S(\mathbf{0}), \frac{\partial}{\partial t_j}S(\mathbf{0})\right) = \mathbf{I}_{\{i=j\}} \int \left[\frac{\partial}{\partial x_i}f(\mathbf{x})\right]^2 e^{\theta f(\mathbf{x})}d\mathbf{x}, \\ &\mathrm{Cov}_\theta(S(\mathbf{0}), \nabla^2 S(\mathbf{0})) = \int f(\mathbf{x})\nabla^2 f(\mathbf{x})e^{\theta f(\mathbf{x})}d\mathbf{x}, \end{aligned}$$

they obtained the approximation

$$E_0 K \sim \theta^{d-1}e^{-I(\theta)}(2\pi)^{-(d+1)/2}|D| \left\{ \frac{\prod_{i=1}^d \mathrm{Var}_\theta\left(\lambda^{1/2}\frac{\partial}{\partial t_i}S(\mathbf{0})\right)}{\mathrm{Var}_\theta(\lambda^{1/2}S(\mathbf{0}))} \right\}^{1/2}.$$

Rabinowitz and Siegmund (1997) also considered scaling of f by an unknown σ to capture clusters of different sizes. Consider the more general score function

$$S(\mathbf{t}, \sigma) = \lambda^{-1/2}\left[\sigma^{-d/2}\sum_{j=1}^{\infty} f\left(\frac{\mathbf{x}_j - \mathbf{t}}{\sigma}\right) - \sigma^{d/2}\lambda\mu\right],$$

and let the scan statistic $M_{\sigma_0,\sigma_1} = \sup_{\mathbf{t}\in D, \sigma_0\le\sigma\le\sigma_1} S(\mathbf{t}, \sigma)$, where $0 < \sigma_0 < \sigma_1 < \infty$. We refer the reader to Rabinowitz and Siegmund (1997) pp.175-179 for the tail approximation of M_{σ_0,σ_1}, which involves a more complicated derivation.

Kernel functions containing discontinuities. When f is not continuous, then $S(\mathbf{t})$ is also not continuous and the approach given above does not work. We illustrate the general approach with the box-shaped kernel

$$f = \mathbf{I}_{A_\Delta}, \text{ where } A_\Delta = \{(x_1, x_2) : 0 \le x_1 \le \Delta_1, 0 \le x_2 \le \Delta_2\},$$

considered in Loader (1991). Let $N(\mathbf{t}, \Delta)$ denote the number of points $\mathbf{x}_j$ lying inside $\mathbf{t} + A_\Delta$. Let $D = [0, 1]^2$ and consider $(\mathbf{t}, \Delta)$ such that $\mathbf{t} + A_\Delta \subset D$. We shall use as our score function at $(\mathbf{t}, \Delta)$, the log GLR test statistic for testing

H_0: intensity of Poisson process is λ at all $\mathbf{t} \in D$,
vs H_1: intensity at $\mathbf{x}$ is $\lambda(\mathbf{x}) = \lambda \exp(\theta \mathbf{I}_{\{\mathbf{x} \in \mathbf{t}+A_\Delta\}})$ for some $\theta > 0$.

Let $\mathbf{t} \prec \mathbf{u}$ if $t_i < u_i$ for all i. Then

$$S(\mathbf{t}, \Delta) = \Big\{ N(\mathbf{t}, \Delta) \log\left(\frac{N(\mathbf{t}, \Delta)}{n\Delta_1\Delta_2}\right) + [n - N(\mathbf{t}, \Delta)] \log\left(\frac{n - N(\mathbf{t}, \Delta)}{n(1 - \Delta_1\Delta_2)}\right) \Big\} \mathbf{I}_{\{N(\mathbf{t},\Delta) \geq n\Delta_1\Delta_2\}}, \quad (1.11)$$

where n is the total number of points in D, and we consider the scan statistic

$$M_{\mathbf{w}_1,\mathbf{w}_2} = \sup_{\mathbf{w}_1 \prec \Delta \prec \mathbf{w}_2} \left[\sup_{\mathbf{t}+A_\Delta \subset D} S(\mathbf{t}, \Delta) \right], \quad (1.12)$$

for some $\mathbf{0} \prec \mathbf{w}_1 \prec \mathbf{w}_2$.

Loader (1991) first considered the case of fixed Δ and n. Let $D' = [0, 1 - \Delta_1] \times [0, 1 - \Delta_2]$ and consider the lattice grid $D'_\delta = D' \cap (\delta\mathbf{Z})^2$. Let $M = \sup_{\mathbf{t} \in D'} N(\mathbf{t}, \Delta)$ and $M_\delta = \sup_{\mathbf{t} \in D'_\delta} N(\mathbf{t}, \Delta)$. Let $P^{(n)}$ denote probability conditioned on n. Using the first-passage time approach given in (I) of Section 2, the tail approximations of $M_\delta := \sup_{\mathbf{t} \in D'_\delta} N(\mathbf{t}, \Delta)$ is first obtained. By using a good bound of $P^{(n)}\{M - M_\delta > 0\}$ for small $\delta > 0$, Loader (1991) showed that for any $\epsilon > 0$ with $\Delta_1\Delta_2(1 + \epsilon)$ rational,

$$P^{(n)}\{M \geq m\} \sim \frac{n^2\Delta_1\Delta_2(1 - \Delta_1)(1 - \Delta_2)\epsilon^3}{(1 - \Delta_1\Delta_2)^3(1 + \epsilon)} P^{(n)}\{N(\mathbf{0}, \Delta) = m\},$$

as $m \to \infty$ with $m = n\Delta_1\Delta_2(1 + \epsilon)$ a positive integer.

We shall now proceed to the tail probabilities of $M_{\mathbf{w}_1,\mathbf{w}_2}$. For given $\eta > 0$, let $h(\Delta)$ be defined implicitly as a solution to the equation

$$h(\Delta) \log\left(\frac{h(\Delta)}{\Delta}\right) + [1 - h(\Delta)] \log\left(\frac{1 - h(\Delta)}{1 - \Delta}\right) = \frac{\eta^2}{2}, \quad (1.13)$$

subject to the constraint $h(\Delta) > \Delta$. Let $c = \eta\sqrt{n}$. Then by (1.11) and (1.12),

$$\Big\{ M_{\mathbf{w}_1,\mathbf{w}_2} \geq c^2/2 \Big\} = \left\{ \sup_{\mathbf{w}_1 \prec \Delta \prec \mathbf{w}_2} \sup_{\mathbf{t}+A_\Delta \subset D} [N(\mathbf{t}, \Delta) - nh(\Delta_1\Delta_2)] \geq 0 \right\}. \quad (1.14)$$

The local random walk analysis of $S(\mathbf{t}, \Delta)$ involves both a tangent approximation

$$h(\Delta') \doteq h(\Delta) + (\Delta' - \Delta)h'(\Delta)$$

and a decomposition

$$N(\mathbf{t}', \Delta') - N(\mathbf{t}, \Delta) \doteq Z_1(t'_1 - t_1) + Z_2(t'_2 - t_2)$$

$$+Z_3(t_1' - t_1 + \Delta_1' - \Delta_1) + Z_4(t_2' - t_2 + \Delta_2' - \Delta_2),$$

where $Z_1, \ldots, Z_4$ are independent two-sided Poisson processes. Then

$$P^{(n)}\{M_{\mathbf{w}_1,\mathbf{w}_2} \geq c^2/2\} \sim c^7\phi(c)\int_{u_0}^{u_1} \frac{u^2}{\eta^7[h'(u)]^3}\left(h'(u) - \frac{1-h(u)}{1-u}\right)^4 \times \left(\frac{1-h(u)}{1-u} - \frac{h(u)}{u}\right)^3 \left(\frac{-(1+u)\log u - 2(1-u)}{\sqrt{h(u)(1-h(u))}}\right) du, \qquad (1.15)$$

where $u_0 = w_{10}w_{20}$ and $u_1 = w_{11}w_{21}$ are the areas of the smallest and largest windows respectively. A simulation study conducted in Loader (1991) shows (1.15) to be more accurate than the approximation obtained using an asymptotic Gaussian process argument.

1.3.2 Case-control epidemiological studies

In the detection of disease clusters, we have to adjust for the non-homogeneity of the underlying population both in terms of the population density and the distribution of disease risk factors like gender, age or ethnic group. One way to achieve this is through a case-control epidemiological study, see for example Whittemore *et al.* (1987), Cuzick and Edwards (1990), Diggle (1990) and Kulldorff (1997).

Assume we have a dataset of locations of disease cases and a corresponding dataset of locations of healthy controls. We merge the two datasets into one and denote it by $\{(\mathbf{x}_j, X_j) : 1 \leq j \leq J\}$, $\mathbf{x}_j$ denoting the location of the jth subject with $X_j = 1$ if it corresponds to a case and $X_j = 0$ if it corresponds to a control.

We focus here on the model proposed in Diggle (1990) to test if there exists a location risk factor that increases the occurrence rate of cases. Let $\lambda(\mathbf{x})$ be the rate of generating controls at position $\mathbf{x}$ and let $\rho\lambda(\mathbf{x})e^{\theta g(\mathbf{x},\mathbf{t})}$ be the rate of generating cases at position $\mathbf{x}$ with $\theta > 0$ when there is a risk factor at $\mathbf{t}$ and $\theta = 0$ when there is no risk factor. The semi-parametric likelihood is proportional to

$$\prod_{j=1}^{J}\{[\lambda(\mathbf{x}_j)\rho e^{\theta g(\mathbf{x}_j,\mathbf{t})}]^{X_j}[\lambda(\mathbf{x}_j)]^{1-X_j}\}$$

while the conditional likelihood for given $\mathbf{x}_1, \ldots, \mathbf{x}_J$ and $I = \sum_{j=1}^{J} X_i$ is

$$\frac{\prod_{j=1}^{J} e^{X_j\theta g(\mathbf{x}_j,\mathbf{t})}}{\sum_{\alpha\in U}\prod_{j=1}^{J} e^{\mathbf{I}_{\{j\in\alpha\}}\theta g(\mathbf{x}_j,\mathbf{t})}},$$

where U is the class of all $\binom{J}{I}$ subsets of $\{1, \ldots, J\}$ of size I. Let $\widehat{p}_0 = I/J$ and $\bar{g}(\mathbf{t}) = J^{-1}\sum_{j=1}^{J} g(\mathbf{x}_j, \mathbf{t})$. Then the efficient score statistic for testing the

presence of a localized risk factor at $\mathbf{t}$ is

$$T_{\mathbf{t}} = \sum_{j=1}^{J} (X_j - \widehat{p}_0)[g(\mathbf{x}_j, \mathbf{t}) - \bar{g}(\mathbf{t})]. \tag{1.16}$$

Let the normalized score $S(\mathbf{t}) = T_{\mathbf{t}}/\sqrt{\mathrm{Var}(T_{\mathbf{t}})}$, where $\mathrm{Var}(T_{\mathbf{t}}) = \widehat{p}_0(1-\widehat{p}_0)(J-2)\sum_{j=1}^{J}[g(\mathbf{x}_j,\mathbf{t}) - \bar{g}(\mathbf{t})]^2/(J-1)$. Rabinowitz (1994) obtained p-value estimates of $M = \sup_{t\in D} S(\mathbf{t})$ by applying the tail probability approximation of a Gaussian process having the same covariance structure as $S(\mathbf{t})$. Let $\sigma_{\mathbf{t},\mathbf{u}} = \mathrm{Cov}(S(\mathbf{t}), S(\mathbf{u}))$, $\Lambda_{\mathbf{t}}$ a matrix with (i,j)th element $-\left(\frac{\partial^2\sigma(\mathbf{s},\mathbf{u})}{\partial s_i \partial s_j}\right)\Big|_{\mathbf{s}=\mathbf{u}}$ and $\Lambda'_{\mathbf{t}} = P_{\mathbf{t}}^T \Lambda_{\mathbf{t}} P_{\mathbf{t}}$, where $P_{\mathbf{t}}$ is a $d \times (d-1)$ matrix comprising of orthonormal vectors of the tangent space of the boundary ∂D at $\mathbf{t}$. Then by Knowles and Siegmund (1989) Corollary 2,

$$\begin{aligned} P\{M > b\} \;\approx\; & (2\pi)^{-d/2} b^{d-1} \varphi(b) \left(\int_D |\Lambda_{\mathbf{t}}|^{1/2} d\mathbf{t} \right. \\ & \left. + (\pi/2)^{1/2} b^{-1} \int_{\partial D} |\Lambda'_{\mathbf{t}}|^{1/2} d\mathbf{t} \right). \end{aligned} \tag{1.17}$$

The SaTScan software developed by Kulldorff (2006) and the Information Management Inc. considers $g(\mathbf{x},\mathbf{t}) = \mathbf{I}_{\{\mathbf{x}-\mathbf{t}\|\le w\}}$ for some $w > 0$. Let $m_{\mathbf{t},w}$ and $n_{\mathbf{t},w}$ be the total number of cases and the total number of occurrences (=cases+controls) respectively in $\{\mathbf{u} : \|\mathbf{u}-\mathbf{t}\| \le w\}$. Instead of the efficient score statistic, they consider the log GLR score

$$S(\mathbf{t}, w) = [n_{\mathbf{t},w}\phi(m_{\mathbf{t},w}/n_{\mathbf{t},w}) + (I - n_{\mathbf{t},w})\phi((J-m_{\mathbf{t},w})/(I-n_{\mathbf{t},w}))]\mathbf{I}_{\{m_{\mathbf{t},w}/n_{\mathbf{t},w} > \hat{p}_0\}},$$

where $\phi(p) = p\log(p/\widehat{p}_0) + (1-p)\log[(1-p)/(1-\widehat{p}_0)]$. In the SaTScan software, p-values of the scan statistics, including scan statistics involving other types of data, are computed using permutation tests.

1.4 Recent Applications in Genomics

Scan statistics are useful for interpreting genomes in the post-sequencing phase. They play an exploratory role, with the goal of locating genomic regions exhibiting properties of extreme deviation to be singled out for further testing. There is a rich source of statistical problems here, many still relatively unexplored. Due to space constraints, we focus only on two examples because the description and solution of each category of problems require a different set of domain knowledge. The first problem is on the scanning of a DNA sequence for predefined word patterns and the second on the analysis of genomic profiling data, in particular DNA copy number profiling.

1.4.1 Biomolecular sequence analysis

DNA and protein sequences can be modeled as a linear sequence drawn from a stationary distribution on an alphabet representing either the 21 amino-acids in the case of protein sequences, or the bases A,C,G and T in the case of DNA sequences. Over the years, researchers have identified specific word patterns that are associated with either the encouragement or suppression of certain biological activity.

Transcription factors are proteins that bind to specific parts of DNA, known as transcription factor binding sites (TFBS), to control the timing and rate of transcription of DNA into RNA. The TFBS are identified by scoring with respect to certain scoring matrices and the presence of a cluster of these sites indicates that genes regulated by the associated transcription factors may be located nearby. Lifanov *et al.* (2003) successfully used scan statistics to locate clusters of binding sites in DNA sequences by counting the number of TFBS located in a sliding window while Rajewsky *et al.* (2002) weighs the TFBS by the scores obtained from the scoring matrices.

A more classical application of scan statistics in counting word patterns is in the identification of origins of replication in viruses, cf. Masse *et al.* (1992). The four letters in the DNA alphabet can be divided into two complementary pairs with A–T one pair and C–G the second pair. In DNA sequences, a palindrome is a DNA word which, when read backwards, has the complementary spelling of the original word. For example, the word ACGCGCGT is a palindrome because its letter-wise complementary spelling is TGCGCGCA. In bacterial and viral genomes, palindromes occur with unusually high frequency near locations associated with the initiation of replication, known as origins of replication.

Karlin and Brendel (1992) formulated the r-scan statistic to detect anomalies in the spacing between occurrences of word patterns. Let n be the length of the genomic sequence and $x_1 < \cdots < x_J$ the locations of the patterns. Let $d_j = x_{j+1} - x_j$ be the inter feature distances, $A_i^{(r)} = \sum_{k=i}^{i+r-1} d_k$ the r-scan process and $A^{(r)} = \min_{1\le i\le J-r} A_i^{(r)}$ the minimal r-scan. Let $N_u(t)$ be the number of word patterns in the interval $(t, t+u]$ and $M_u = \sup_{0\le t\le n-u} N_u(t)$ the maximal scan statistic. Then we have the duality

$$\{M_u \ge r+1\} = \{A^{(r)} \le u\},$$

and the two scan statistics can be used interchangeably. P-value approximations for the significance of r-scans were obtained by Arratia, Goldstein, and Gordon (1989) and Glaz *et al.* (1994) using Poisson and compound Poisson approximations respectively, see also Leung and Yamashita (1999) for the applications of these p-value approximations on palindrome counting scan statistics.

In addition to Rajewsky *et al.* (2002), weighted scan statistics was also considered in Chew, Choi and Leung (2005) for scoring palindromic patterns,

which we consider here to be palindromes having length of at least ten DNA letters. Since the length of a palindrome must be even, they let $X_j = \ell_j/2$, where ℓ_j is the length of the jth palindromic pattern. Let $S_u(t) = \sum_{\mathbf{x}_j \in (t,t+u]} X_j$ and let the weighted scan statistic $M_{n,u} = \sup_{0 \le t \le n-u} S_u(t)$. Chan and Zhang (2007) used a marked Poisson process approximation of $S_u(t)$ to obtain an approximation of the p-value of $M_{n,u}$. Let F be the distribution of X_j, which we assume to have positive mean μ. Let λ be the probability of observing a palindromic pattern at any one location. Let $K(\theta) = E(e^{\theta X_1})$ and for given $x > \lambda\mu$, define $\theta_x(> 0)$ and $\alpha_x(> \lambda)$ to be the unique constants satisfying

$$K'(\theta_x) = x/\lambda, \qquad \alpha_x = \lambda K(\theta_x). \tag{1.18}$$

Let the large deviation rate function $I(x) = -(\alpha_x - \lambda) + \theta_x x$ and define F_θ to be the tilted distribution of F satisfying $F_\theta(dx) = e^{\theta x} F(dx)/K(\theta)$, with probability mass function (density) f_θ when F is discrete (continuous). Let $Y_1, Y_2, \ldots$ be i.i.d. random variables with the mixture probability mass function (density)

$$g(y) = \Big(\frac{\alpha_x}{\lambda + \alpha_x}\Big) f_{\theta_x}(y) + \Big(\frac{\lambda}{\lambda + \alpha_x}\Big) f(-y), \tag{1.19}$$

and let $R_k = Y_1 + \cdots + Y_k$. Define the overshoot constant

$$\nu_x = \lim_{b \to \infty} E[e^{-\theta_x (R_{\tau_b} - b)}], \text{ where } \tau_b = \inf\{k \ge 1 : R_k \ge b\}, \tag{1.20}$$

with b a multiple of η if F is arithmetic with span η, in other words, if F has support on the grid $\{0, \pm\eta, \pm 2\eta, \ldots\}$ but not on a coarser lattice grid containing 0. By the approach of conditioning below a high crossing, see (III) in Section 2, Chan and Zhang (2007) showed that

$$P\{M_{n,u} \ge ux\} \sim 1 - \exp\left\{-\frac{(n-u)\nu_x e^{-uI(x)}(x - \lambda\mu)}{\sqrt{2\pi u \lambda K''(\theta_x)}}\right\}, \tag{1.21}$$

if $u \to \infty$ and $(n-u) \to \infty$ as $n \to \infty$.

In Figure 1.1, we use (1.21) to obtain threshold levels corresponding to a p-value of 0.001 in the search for clusters of palindromic patterns with window size u equal to 0.5 % of the genome length. For the unweighted case, $X_j = 1$ for all palindromic patterns, while for the weighted case, we choose $X_j = (\ell_j/2) - 4$.

1.4.2 Detecting changes in DNA copy number

DNA copy number is the number of copies of DNA at a region of a genome, the default being two for all human autosomes. The variation of this number, known as DNA copy number variation (CNV), corresponds to gains and losses of specific chromosomal segments. These variations may be inherited [Redon *et al.* (2006)], or they may occur due to mutation and are then associated

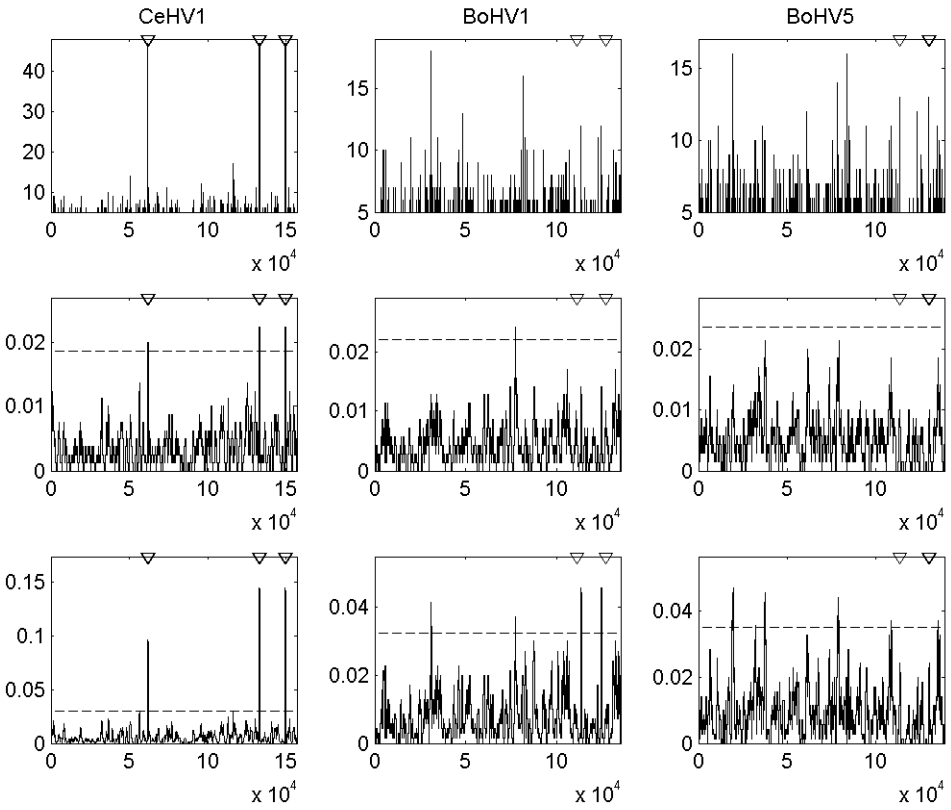

Figure 1.1: The x co-ordinate represents the locations of three well-known virus genomes. The y co-ordinate represents either half the length of the palindromic patterns (top plots), $u^{-1}N_u(t-u/2)$ for the unweighted case (middle plots) or $u^{-1}S_u(t-u/2)$ for the weighted case (bottom plots). The dotted lines are threshold levels corresponding to p-values of 0.001. The inverted triangles are experimentally validated origins of replication.

with certain diseases like cancer [Pinkel and Albertson (2005)]. In DNA copy number data, the quantity of homologous DNA present in a population of cells is measured by a set of probes, each mapping to a specific location in the genome.

Let X_j be the measured DNA quantity at probe j, relative to the expected value of two, at a fixed location x_j in the genome. We do not observe integer valued X_j due to inhomogeneity of the cell sample and substantial measurement error. Our objective is to partition the genome into segments of equal copy number. We shall disregard irregularities in the spacing of the probe locations, a reasonable assumption for most experimental platforms and accepted in practice. Many different statistical methods have been applied to this problem, see Lai *et al.* (2005) for a broad survey of these methods. We shall focus here on the approach taken by Olshen *et al.* (2004). Consider a segment of the genome, containing J probes, which we would like to test for constant CNV. Define $\bar{X} = J^{-1}\sum_1^J X_j$ and $\hat{\sigma}^2 = J^{-1}\sum_1^J (X_j - \bar{X})^2$. Let

$$U(s,t) = \frac{\sum_{j=s+1}^{t}(X_j - \bar{X})}{\hat{\sigma}\sqrt{(t-s)[1-(t-s)/J]}}, \tag{1.22}$$

and

$$M = \max_{0\le s<t\le J, v_0<t-s<v_1} U^2(s,t). \tag{1.23}$$

When a significant p-value is obtained, for example by using the approximation in Siegmund (1986), we partition the segment and test each sub-segment further in the same manner.

Since most genomic profiling studies involve cohorts of individuals, it is of interest to pool samples together to gain power in detecting recurrent CNVs. This problem was first analyzed using hidden Markov models, cf. Shah *et al.* (2007), and has also been studied recently by Zhang *et al.* (2008) under the framework of a simultaneous scan of multiple aligned sequences for recurrent variant intervals of shared location. The formulation in Zhang *et al.* (2008) is as follows. For each sequence $i = 1, \ldots, N$ and position $j = 1, \ldots, J$, the random variables X_{ij} are mutually independent and normally distributed with mean values μ_{ij} and variances σ_i^2. Under the null hypothesis, $\mu_{i1} = \cdots = \mu_{iJ}$ for each sample i and under the alternative hypothesis, there exists $\mathcal{J} \subset \{1, \ldots, N\}$ (with $\mathcal{J} \neq \emptyset$), and $\tau_1 < \tau_2$ with $v_0 \le (\tau_2 - \tau_1) \le v_1$ for some $1 \le v_0 \le v_1 < J$, such that for each $i \in \mathcal{J}$, $\mu_{ij} = \mu_{i0} + \delta_i \mathbf{I}_{\{\tau_1 < j \le \tau_2\}}$ for some $\delta_i \neq 0$. The GLR test in this setting yields the scan statistic

$$M = \max_{0\le s<t\le J, v_0\le t-s\le v_1} Z_{s,t}, \quad \text{where } Z_{s,t} = \sum_{i=1}^{N} \frac{[U_i^2(s,t) - 1]}{\sqrt{2N}}, \tag{1.24}$$

and $U_i(s,t)$ is defined as in (1.22) relative to the ith sequence.

The sum of chi-squares statistic in (1.24) pools signals from all samples, however weak. Zhang *et al.* (2008) also proposed a weighted sum of chi-squares statistic that requires individual sequences to show some evidence of a signal before it is allowed to contribute significantly to the pooled scan. Let $Q_i(s,t) = \mathbf{I}_{\{i\in\mathcal{J}\}}$ (the presence of (s,t) in the notation will be clear later on). If $\mathcal{J}$ is known, then the log likelihood ratio statistic is

$$\max_{s<t}\sum_{i=1}^{N}\log\{[1-Q_i(s,t)]+Q_i(s,t)e^{U_i^2(s,t)/2}\} = \max_{s<t}\sum_{i=1}^{N}Q_i(s,t)U_i^2(s,t)/2. \tag{1.25}$$

Since $Q_i(s,t)$ is not observable, a plug-in estimate is derived by using a Bayesian formulation. Let p denote the prior probability that $Q_i(s,t)=1$. Then the posterior mean of $Q_i(s,t)$, after maximizing with respect to the unknown parameters, is

$$\widehat{Q}_i(s,t) = \frac{e^{U_i^2(s,t)/2}}{r_p + e^{U_i^2(s,t)/2}}, \tag{1.26}$$

where $r_p = (1-p)/p$. Replacing Q_i by $\widehat{Q}_i$ in (1.25) and standardizing leads to the weighted sum of chi-squares statistic

$$Z^{(p)}(s,t) = \frac{\sum_{i=1}^{N}[w(U_i(s,t))U_i^2(s,t)-\mu_p]}{\sigma_p\sqrt{N}}, \tag{1.27}$$

where $w(u) = e^{u^2/2}/\{r_p + e^{u^2/2}\}$ and μ_p, σ_p^2 are the mean and variance respectively of $w(U)U^2$ when U is a standard normal random variable.

An approximation of the significance of scans using either (1.24) or (1.27) can be obtained via a last exit-time approach. Instead of the process $Z_{s,t}^{(p)}$, we consider more generally

$$Z_{s,t}^{f} = \frac{\sum_{i=1}^{N}[f(U_i(s,t))-\mu]}{\sigma\sqrt{N}},$$

where f is a well-behaved function, $\mu = Ef(U)$ and $\sigma^2 = \mathrm{Var}(f(U))$. Under the assumption that the noise is independent between samples, $Z_{s,t}^{f}$ is a normalized sum of N i.i.d. processes, and thus for large N is approximately a mean zero Gaussian process on the two dimensional indexing set $D = \{(s,t) : 0 \le s < t \le J, v_0 \le t-s \le v_1\}$ with covariance function

$$\rho(s,t,u,v) = \mathrm{Cov}(Z_{s,t}^{f}, Z_{u,v}^{f}) = \sigma^{-2}\mathrm{Cov}(f(U_1(s,t)), f(U_1(u,v))). \tag{1.28}$$

The function ρ is not differentiable, but its left and right partial derivatives exist and have the same magnitude. Hence we may define

$$\rho'(s,t) = \lim_{a\uparrow 0}\left|\frac{\rho(s,t,s+a,t)-\rho(s,t,s,t)}{a}\right|. \tag{1.29}$$

By conditioning on the last exit-time, it follows from the calculations in Siegmund (1988) that

$$P\left\{\max_{(s,t)\in D} Z^f_{s,t} > c\right\} \approx \frac{\varphi(c)}{c}\sum_{(s,t)\in D}\int_0^\infty e^{-x} P\left\{\max_{n\geq 1} W_n^{(s,t)} \leq -x\right\} \times P\left\{\min_{n\geq 0} W_n^{(s,t)} + \min_{n\geq 1}\widetilde{W}_n^{(s,t)} \geq x\right\} dx, \quad (1.30)$$

where $W_n^{(s,t)}$ is a random walk of i.i.d. normal random variables with mean $-c^2\rho'(s,t)$ and variance $2c^2\rho'(s,t)$, and $\widetilde{W}_n^{(s,t)}$ is an identically distributed random walk, independent of the first random walk. The formula in (1.30) uses a Gaussian approximation on $Z^f_{s,t}$, which is asymptotically a function of the chi-square random variables.

A more accurate approximation can be obtained by correcting for the skewness of $f(U)$. Let $\psi(\theta) = \log\exp\{\theta[f(U)-\mu]/\sigma\}$ and select θ to be the positive root of the equation $N^{1/2}\psi'(\theta) = c$. Replace $\varphi(c)/c$ in (1.30) with the saddlepoint approximation $[2\pi\psi''(\theta)]^{-1/2}\exp\{-N[\theta\psi'(\theta)-\psi(\theta)]\}$ and use Lemma 21 of Siegmund (1992) to evaluate the integral to obtain

$$P\left\{\max_{(s,t)\in D} Z^f_{s,t} > c\right\} \approx [2\pi\psi''(\theta)]^{-1/2} e^{-N[\theta\psi'(\theta)-\psi(\theta)]} c^3 \times \sum_{(s,t)\in D}[\rho'(s,t)]^2\nu^2\left(c_0[2\rho_1'(s,t)]^{1/2}\right), \quad (1.31)$$

where $c_0 = c/\sqrt{N}$ and ν is the overshoot constant given in (1.6).

The computation of the partial derivatives ρ' can be simplified by using the expression

$$\rho'(s,t) = (2\sigma^2)^{-1}\{E[f(U_{s,t})f'(U_{s,t})U_{s,t}] - E[f(U_{s,t})f''(U_{s,t})]\}\kappa(t-s), \quad (1.32)$$

where $\kappa(r) = [r(1-r/J)]^{-1}$. For example, $f(x) = x$ corresponds to the simple one sample case and by (1.32), $\rho'(s,t) = \kappa(t-s)/2$. Plugging this inside (1.31) provides us with the significance level approximation of Siegmund (1992). The sum of chi-squares statistic (1.24) corresponds to $f(x) = x^2$ and $\rho'(s,t) = \kappa(t-s)$.

1.5 Concluding remarks

In addition to DNA copy number, scan statistics can be applied on many other types of genomic profiling data. Recent technological advancements have allowed the measurement of many types of genomic activity, all of which produce

enormous quantities of data where the primary goal is to locate regions of change from baseline in a linear sequence. There is a common theme of scanning for signals of unknown width and scanning for simultaneous signals in multiple sequences. Hoh and Ott (2000), Ji and Wong (2005) and Keles, van der Laan, Dudoit and Cawley (2006) are recent articles that applies scan statistics on the DNA genome. These advancements and advancements in other applied fields like neuroscience, have resulted in the collection of a huge amount of data and scan statistics have been useful in identifying meaningful signals and patterns.

In more traditional areas of scan statistics applicatons, for example in astronomy and epidemiological studies, there are still many important issues that can occupy the attention and time of researchers. Current scan statistics are geared towards the detection of one cluster of a pre-determined shape. It will be interesting to study how scan statistics can be modified so that they can detect multiple clusters or signals with irregular shapes more effectively.

References

1. Adler, R.J. (1981). *The Geometry of Random Fields*, Wiley, New York.

2. Aldous, D. (1989). *Probability Approximations via the Poisson Clumping Heuristic*, Springer-Verlag, New York.

3. Arratia, R., Goldstein, L., and Gordon, L. (1989). Two moments suffice for Poisson approximations: The Chen-Stein method, *Annals of Probability*, **17**, 9–25.

4. Bickel, P. and Rosenblatt, M. (1973). Two-dimensional random fields, In *Multivariate Analysis–III* (Ed., P.K. Krishnaiah), pp.3–15, Academic Press, New York.

5. Chan, H.P. and Lai, T.L. (2002). Boundary crossing probabilities for scan statistics and their applications to change-point detection, *Methodology and Computing in Applied Probability*, **4**, 317–336.

6. Chan, H.P. and Lai, T.L. (2003). Saddlepoint approximations and nonlinear boundary crossing probabilities of Markov random walks, *Annals of Applied Probability*, **13**, 395–429.

7. Chan, H.P. and Loh, W.L. (2007). Some theoretical results on neural spike train probability models, *Annals of Statistics*, **35**, 2691–2722.

8. Chan, H.P. and Zhang, N.R. (2007) Scan statistics with weighted observations, *Journal of the American Statistical Association*, **102**, 595–602.

9. Chew, D., Choi, K. and Leung, M. (2005). Scoring schemes of palindrome clusters for more sensitive prediction of replication origins in herpesviruses, *Nucleic Acids Research*, **33**, e134.

10. Chi, Z. (2004). Large deviations for template matching between point processes, *Annals of Applied Probability*, **15**, 153–174.

11. Chi, Z., Rauske, P.L. and Margoliasch, D. (2003). Pattern filtering for detection of neural activity, with example from HVc activity during sleep in zebra finches, *Neural Computing*, **15**, 2307–2337.

12. Cuzick, J. and Edwards, R. (1990). Spatial clustering for inhomogeneous populations, *Journal of the Royal Statistical Society, Series B*, **52**, 73-104.

13. Dave, A.S. and Margoliasch, D. (2000). Song replay during sleep and computational rules for sensorimotor vocal learning, *Science*, **290**, 812–816.

14. Diggle, P.J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point. *Journal of the Royal Statistical Society, Series A*, **153**, 349–362.

15. Glaz, J. (1989). Approximations and bounds for the distribution of the scan statistic, *Journal of the American Statistical Association*, **84**, 560–566.

16. Glaz, J., Naus, J., Roos, M. and Wallenstein, S. (1994). Poisson approximations for the distribution and moments of ordered m-spacings. *Journal of Applied Probability*, **31**, 271–281.

17. Glaz, J. Naus, J. and Wallenstein, S. (2001). *Scan Statistics*, Springer-Verlag, New York.

18. Hogan, M. L. and Siegmund, D. (1986). Large deviations for the maxima of some random fields, *Advances in Applied Mathematics*, **7**, 2–22.

19. Hoh, J. and Ott, J. (2000). Scan statistics to scan markers for susceptibility genes, *Proceedings of the National Academy of Sciences*, **97**, 9615–9617.

20. Huntington, R. and Naus, J. (1975). A simple expression for kth nearest neighbor coincidence probabilities, *Annals of Probability*, **3**, 894–896.

21. Ji, H. and Wong, W.H. (2005), TileMap: create chromosomal map of tiling array hybridizations, *Bioinformatics*, **21**, 3629–3636.

22. Karlin, S. and Brendel, V. (1992). Chance and statistical significance in protein and DNA sequence analysis, *Science*, **257**, 39–49.

23. Keles, S., van der Laan, M., Dudoit, S. and Cawley, S.E. (2006). Multiple testing methods for ChIP-Chip high density oligonucleotide array data, *Journal of Computational Biology*, **13**, 579–613.

24. Knowles, M. and Siegmund, D. (1989). On Hotelling's approach to testing for a nonlinear parameter in regression, *International Statistical Review*, **57**, 205–220.

25. Kulldorff, M. (1997). A spatial scan statistic, *Communications in Statistics: Theory and Methods*, **26**, 1481–1496.

26. Kulldorff, M. (2006). *SaTScan User Guide*, http://www.satscan.org/techdoc.html.

27. Lai, T.L. and Siegmund, D. (1977, 1979). A nonlinear renewal theorem with applications to sequential analysis I, *Annals of Statistics*, **5**, 946–955, II, *Annals of Statistics*, **7** 60–76.

28. Lai, W.R., Johnson, M.D., Kucherlapati, R., and Park, P.J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data, *Bioinformatics*, **21**, 3763–3770.

29. Leung, M.Y. and and Yamashita, T.E. (1999). Applications of the scan statistic in DNA sequence analysis, In *Scan Statistics and Applications*. (Ed., J. Glaz and N. Balakrishnan), pp. 269–286, Birkhäuser, Boston.

30. Lifanov, A., Makeev, V., Nazina, A. and Papatsenko, D. (2003). Homotypic regulatory clusters in Drosophila, *Genome Research*, **13**, 579–588.

31. Loader, C. (1991). Large-deviation approximations to the distribution of scan statistics, *Advances in Applied Probability*, **23**, 751–771.

32. Masse, M.J.O., Karlin, S., Schachtel, G.A. and Mocarski, E.S. (1992). Human Cytomegalovirus origin of DNA replication (oriLyt) residues with a highly complex repetitive region, *Proceedings of the National Academy of Science*, **89**, 5246–5250.

33. Naus, J. (1965). Clustering of random points in two dimensions, *Biometrika*, **52**, 263-267.

34. Naus, J. (1966). Some probabilities, expectations, and variances for the size of largest clusters and smallest intervals, *Journal of the American Statistical Association*, **61**, 1191–1199.

35. Naus, J. (1982). Applications for distributions of scan statistics, *Journal of the American Statistical Association*, **77** 177–183.

36. Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data, *Biostatistics*, **5**, 557–572.

37. Pickands, J. (1969). Upcrossing probabilities for stationary Gaussian processes, *Transactions of the American Mathematical Society*, **145**, 51–73.

38. Pinkel, D. and Albertson, D.G. (2005). Array comparative genomic hybridization and its applications in cancer, *Nature Genetics*, **37**, Suppl 11–17.

39. Qualls, C. and Watanabe, H. (1973). Asymptotic properties of Gaussian random fields, *Transactions of the American Mathematical Society*, **177**, 155–171.

40. Rabinowitz, D. (1994). Detecting clusters in disease incidence, In *Changepoints Problems* (Ed., E. G. Carlstein, H.-G. Muller and D. Siegmund), 255–275, IMS, Hayward.

41. Rabinowitz, D. and Siegmund, D. (1997). The approximate distribution of the maximum of a smoothed Poisson random field, *Statistica Sinica*, **7**, 167–180.

42. Rajewsky, N., Vergassola, M., Gaul, U. and Siggia, E. (2002). Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early Drosophila embryo, *BMC Bioinformatics*, **3**, e30.

43. Redon, R., Ishikawa, S. Fitch, K.R., Feuk, L., Perry, G.H. *et al.* (2006). Global variation in copy number in the human genome, *Nature*, **444**, 444-454.

44. Shah, S.P., Lam, W.L., Ng, R.T. and Murphy, K.P. (2007). Modeling recurrent DNA copy number alterations in array CGH data, *Bioinformatics*, **23**, 450–458.

45. Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*, Springer-Verlag, New York.

46. Siegmund, D. (1986). Boundary crossing probabilities and statistical applications, *Annals of Statistics*, **14**, 361–404.

47. Siegmund, D. (1988). Tail probabilities for the maxima of some random fields. *Annals of Probability*, **16**, 487–501.

48. Siegmund, D. (1992). Tail approximations for maxima of random fields, In *Probability Theory: Proceedings of the 1989 Singapore Probability Conference* (Ed., L. H. Y. Chen, K. P. Choi, K. Hu and J.-H. Lou), pp. 147–158, Walter de Gruyter, Berlin.

49. Siegmund, D. and Venkatraman, E.S. (1995). Using the generalized likelihood ratio statistic for sequential detection of a change-point, *Annals of Statistics*, **23**, 255–271.

50. Siegmund, D. and Yakir, B. (2000). Tail probabilities for the null distribution of scanning statistics, *Bernoulli*, **6**, 191–213.

51. Storey, J.D. and Siegmund, D. (2001). Approximate p-values for local sequence alignments: numerical studies. *Journal of Computational Biology*, **8**, 549–556.

52. Tu, I. (2008). Asymptotic overshoots for arithmetic i.i.d. random variables, to appear in *Statistica Sinica.*

53. Tu, I. and Siegmund, D. (1999). The maximum of a function of a Markov chain and application to linkage analysis, *Advances in Applied Probability*, **31**, 510–531.

54. Whittemore, A.S., Friend, N., Brown, B. and Holly, E. (1987). A test to detect clusters of diseases, *Biometrika*, **74**, 631–635.

55. Woodroofe, M. (1978). Large deviations of likelihood ratio statistics with applications to sequential testing, *Annals of Statistics*, **6**, 72–84.

56. Woodroofe, M. (1979). Repeated likelihood ratio tests, *Biometrika*, **66**, 454–463.

57. Woodroofe, M. (1982). *Nonlinear Renewal Theory in Sequential Analysis*, SIAM, Philadelphia.

58. Zhang, N.R., Siegmund, D., Ji, H. and Li, J. (2008). Detecting simultaneous change-points in multiple sequences. Technical Report, Department of Statistics, Stanford University.